A Collaborative Human-AI Guided Decision Framework - Gen-Edge-AI

Eman Sayed
Decision Support Department
Faculty of Computers and Informatics
Zagazig University
Zagazig, Egypt
essayed@fci.zu.edu.eg
https://orcid.org/0000-0001-6121-8458

Sara M. Mosaad
Information Systems Department
Faculty of Commerce and Business Administration
Helwan University
Cairo, Egypt
sara.mosaad87@gmail.com
https://orcid.org/0009-0003-4597-6445

Ahmad N. Gohar
Technology and Innovation CoC
IBM
CA, USA
ansgohar@gmail.com
https://orcid.org/0000-0003-2087-1092

2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) | 979-8-3503-6777-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/MIUCC62295.2024.10783537

***Abstract*—Generative AI (Gen-AI) is a cloud-based AI capable of generating new content with applications across numerous fields. This paper analyzes a dataset of 960 Gen-AI tools across 29 domains. To address challenges associated with Gen-AI, such as latency and data privacy, Edge-AI is employed to process data locally. This paper introduces Gen-Edge-AI, a novel human-AI guided decision-making framework that integrates Gen-AI, Edge-AI, and human expertise. By combining technology with human judgment, Gen-Edge-AI provides innovative and ethically grounded solutions. The proposed framework optimally balances the use of Gen-AI and Edge-AI through a decision matrix of routing factors in its Evaluation component. To ensure secure and private decision-making, end-to-end encryption and data masking are employed. Gen-Edge-AI has potential benefits and prospective applications in scenarios requiring human judgment and ethical oversight, including healthcare, law enforcement, and beyond.**

Keywords—Gen-AI, Edge-AI, ethical-AI, human-AI, decision making.

I. Introduction

Although AI systems can significantly enhance decision-making processes, human oversight is indispensable, particularly in contexts requiring ethical judgment and complex decision-making. Human-AI collaboration has demonstrated significant strategic, financial, and environmental advantages, delivering high-quality results in fields such as risk assessment and preventive maintenance [1]. This paper presents key findings, challenges, and advantages of Generative AI (Gen-AI) and Edge-AI. A new human-AI guided decision-making framework, Gen-Edge-AI, is introduced, which combines Gen-AI and Edge-AI with human expertise to provide real-time, intelligent solutions for ethical and high-stakes situations. The human factor is crucial in the proposed model, ensuring that AI-driven decisions remain ethically grounded and contextually appropriate, particularly in critical scenarios where human judgment is essential. This integrated approach enhances system efficiency, responsiveness, and data security, offering innovative solutions with low latency, high security, and optimized resource use. Secure communication protocols, such as end-to-end encryption and data masking, protect sensitive information where necessary. An analysis of 960 Gen-AI tools across 29 fields indicates promising prospects for Gen-Edge-AI, with Millennials and Generation Z expected to be the primary users [2].

The structure of this paper is as follows: Section II discusses the establishment of Gen-AI. Section III explores the application fields of Gen-AI and includes relevant data insights. Section IV provides information on Edge-AI, its development, and its promising applications in various industries. The newly proposed Gen-Edge-AI framework is discussed in Section V. The detailed components of Gen-Edge-AI are discussed in Section VI. Section VII represents the routing factors of Gen-Edge-AI while its potential benefits are discussed in Section VIII. Section IX presents the prospective applications for the proposed framework. Finally, Section X discusses conclusion and future work.

II. Gen-AI

Generative AI (Gen-AI) refers to AI systems capable of creating new content, such as images, text, or video. Popular tools like ChatGPT, Midjourney, Stable Diffusion, and LLaMA fall under this category. Gen-AI represents a unique subset of AI, defined by its ability to produce novel outputs using advanced techniques. Theis field experienced a major breakthrough in 2014 with the introduction of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), along with the development of sequence-to-sequence (seq2seq) architectures. These advancements have driven the use of Gen-AI in creative applications, content generation, and chatbots.

VAEs are commonly applied in data generation, augmentation, and anomaly detection. They are capable of producing text, audio, and images. GANs, on the other hand, consist of two neural networks: a generator, which creates images from random noise, and a discriminator, which differentiates between real and synthetic images. The generator's goal is to deceive the discriminator, while the discriminator aims to correctly classify the images. GANs and VAEs are employed for data augmentation, drug discovery, and various image processing tasks, such as upscaling, inpainting, and colorizing black-and-white photos. Notably, GANs power applications like the model used in "This Person Does Not Exist" website, known for generating highly realistic images [3].

The seq2seq architecture is integral to many Natural Language Processing (NLP) tasks, particularly when combined with attention mechanisms. This setup allows the model to focus on relevant parts of the input when generating output, making it effective for applications like chatbots, machine translation (e.g., Google Translate), and text summarization. Looking ahead, Gen-AI models may evolve to create complete movies, immersive gaming environments, and even metaverses.

The Generative Pre-trained Transformer (GPT) model, a cornerstone of Gen-AI, excels in generating text by leveraging patterns learned during pre-training on diverse datasets. It

979-8-3503-6777-5/24/$31.00 ©2024 IEEE

utilizes a transformer architecture that is well-suited for NLP tasks. Much of modern AI, including applications like speech recognition, image recognition, and spam filtering, relies on the underlying power of deep neural networks.

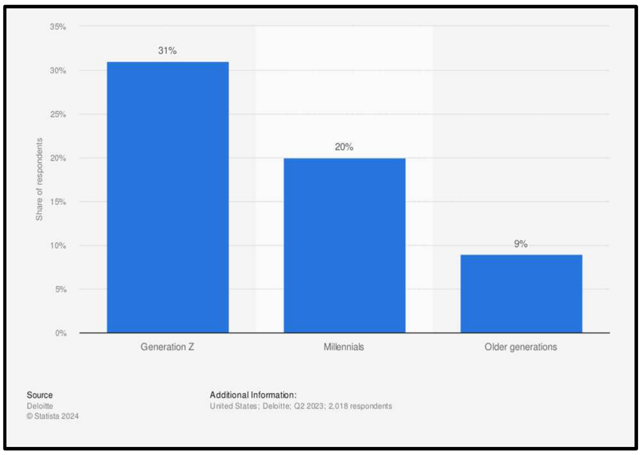
The rapid advancements and transformative potential of Gen-AI have led to substantial investments across various sectors. A survey conducted in the second quarter of 2023 in the United States with 2,018 respondents in Fig. 1 indicates that 31% of Generation Z and 20% of Millennials have used Gen-AI [2]. Furthermore, the International Data Corporation (IDC) forecasts that enterprise spending on Gen-AI software, infrastructure hardware, and related IT/business services will reach $151.1 billion by 2027, with a compound annual growth rate of 86.1% from 2023 to 202 [4].

III. Gen-AI Applications

Gen-AI encompasses a wide range of models, with text and image generators being among the most prominent and widely used. However, the distinction between these two categories is increasingly blurred by the emergence of Large Multimodal Models (LMMs) like Flamingo and Gemini, which can process and generate various data types, including text, audio, and video. Moreover, Gen-AI is finding new applications in creative fields such as music composition and video generation. Despite these advancements, many of these applications still heavily rely on text- or image-based generative models or a combination of both. An analysis in [5] identified seven key clusters of research topics in Gen-AI: image processing and content analysis, content generation, emerging use cases, engineering, cognitive inference and planning, data privacy and security, and GPT academic applications.

Gen-AI has been applied across diverse fields, such as teaching and learning [6], addressing real-world challenges in the public sector [7], mask detection and social distancing [8], and cancer care [9]. This paper analyzes a dataset of 960 Gen-AI tools [10], which serves as a valuable resource for individuals and organizations seeking to explore available tools and identify those best suited to their specific needs.

The analysis of this dataset offers insights into the AI fields with the most developed tools, as illustrated in Fig. 2, which shows the percentage of tools in each category. The Productivity and Copyright fields have a similar number of developed tools, while the Generative Art and Chat categories also show comparable shares of Gen-AI tools. The "Other" category includes fields such as Finance, Speech-to-Text, AI Detection, Gaming, Inspiration, Podcasting, Voice Modulation, Motion Capture, Image Scanning, and Text-to-Video.



Source
Deloitte
© Statista 2024

Additional Information:
United States; Deloitte; Q2 2023; 2,018 respondents

Fig. 1. Users' generation who tried Gen-AI as at Q2, 2023.

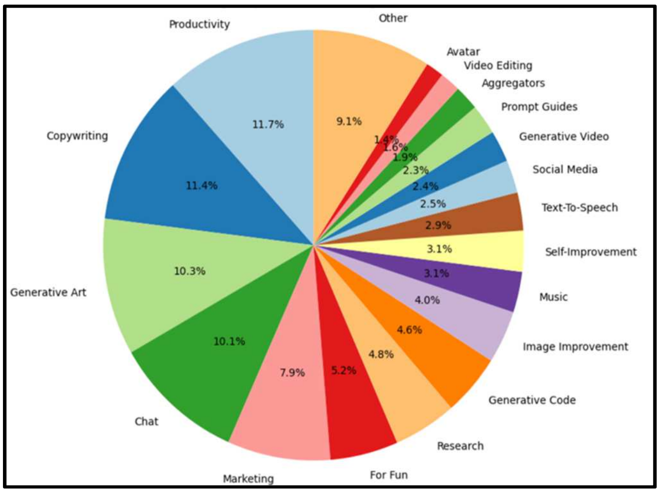


Fig. 2. Gen-AI tools Categories.

Fig. 3 highlights the top three tools in the five highest-ranked application fields. Upvotes represent user approval for AI tools, indicated by interactions with the tool's source or provider. Many of these categories require real-time responses with minimal latency, and some demand enhanced security and privacy measures. This underscores the importance of Edge-AI models in addressing these challenges, as discussed in the next section.

IV. Edge-AI

Edge-AI refers to the deployment of AI models on edge devices that are located near the data source at the periphery of the network, relying minimally on cloud computation [4]. Unlike Gen-AI, which often relies on cloud infrastructure, Edge-AI processes data locally on devices, such as smartphones or IoT equipment. Edge computing brings computation closer to the data source, reducing the need to transmit data to centralized cloud servers. It functions like a mini data center located directly where the action takes place. According to the International Data Corporation (IDC), global investment in edge computing is projected to increase by 15.4% by the end of 2024 compared to 2023. Additionally, this investment is expected to grow significantly by 2027, representing nearly a 50.86% increase from 2024 [4].

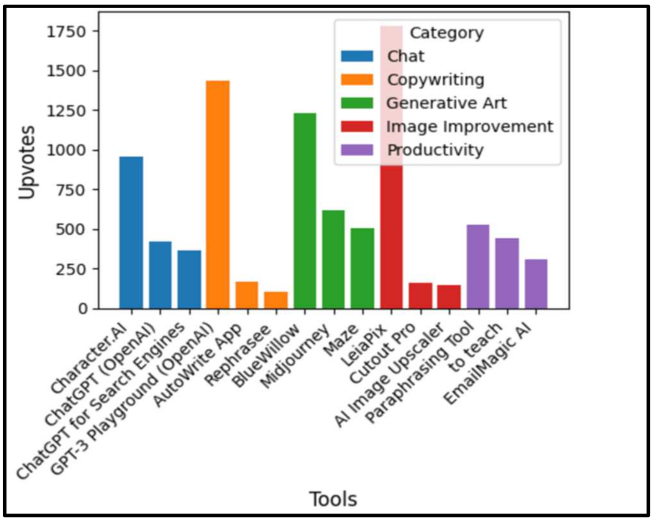


Fig. 3. Top 3 voted Gen-AI tools in the top 5 categories.

Techniques like quantization, pruning, and knowledge distillation are used to compress AI models into smaller, more efficient versions [11]. These "mini models" can perform local processing with high accuracy. For instance, the TinyLlama model [12], launched in 2024 after extensive training, occupies less than 640 megabytes of storage. It was trained on a trillion tokens and outperforms similar-sized models, bringing the capabilities of large-scale models into a compact, mobile-friendly package suitable for edge devices. Mini models can be deployed on various edge devices, including smartphones, personal computers, drones, IoT devices, autonomous vehicles, smart transportation systems, XR headsets, industrial automation machines, and smart city infrastructure. Another application of Edge-AI on personal devices is smart keyboards, which use AI to learn a user's accent or predict the next word based on typing history. The usability and efficiency of various Edge-AI devices were assessed demonstrating that

By processing data at or near its point of creation, Edge-AI provides significant benefits, such as reduced latency, improved real-time responsiveness for timely decision-making, enhanced data protection [13], and cost-effective inference [14]. Edge-AI models can also be optimized for energy efficiency, even in resource-constrained environments [15]. A literature performance test in [14] was conducted on six edge devices: Raspberry Pi, Nvidia's Jetson Nano, TX2, NX, Rockchip's RK3399Pro, and Bitmain's SE5. In that test, three core deep learning tasks were deployed on each of the six devices: object detection, image classification, and natural language processing. This test demonstrated that Edge-AI devices achieved remarkable gains compared to conventional edge devices, with improvements of 134x in throughput, 57x in power efficiency, and 32x in cost-effectiveness. These significant improvements highlight the value of incorporating Edge-AI into the proposed Gen-Edge-AI framework, which is discussed in the following section.

V. Gen-Edge-AI Methodology

While significant progress has been made in developing advanced Gen-AI models, further improvements are necessary to achieve high-performance outcomes with minimal latency, strong security, and optimized cost efficiency [5]. That encouraged its integrated with Edge-AI to efficiently manage computational demands, minimize power consumption, and support scalable, responsive solutions in edge environments. Gen-AI and Edge-AI are transformative technologies that, when combined with human expertise, can provide real-time decision-making with innovative solutions for requests that are subjective and often need contextual understanding. This paper presents a novel Gen-Edge-AI framework that involve a human expert component which makes this framework highly responsive in high-stakes and ethical situations where human judgment is essential. This framework supports adaptive, reliable, and highly responsive decision-making, while safeguarding data privacy through secure Gen-AI cloud communication measures such as end-to-end encryption and data masking.

A key component of the Gen-Edge-AI framework is the human expert, who provides critical oversight and responsive decision-making to direct the initial request to Gen-AI or Edge-AI. This direction requires the human expert to evaluate factors that are subjective and often need contextual understanding, which is best provided by a human expert. While AI can assist, human authority remains crucial, particularly in scenarios that involve judgment and ethical considerations. Human-AI interaction models have shown superior strategic, financial, and environmental benefits, as well as high-quality outcomes in areas like risk assessment and preventive maintenance [16]. The rapid advancements in Gen-AI and Edge-AI present compelling opportunities for innovative problem-solving when guided by human expertise [17]. This framework co-creates solutions by integrating Gen-AI and Edge-AI, while ensuring human control in critical and ethical situations. The subsequent sections discuss the components of the proposed framework and the critical role of human expertise.

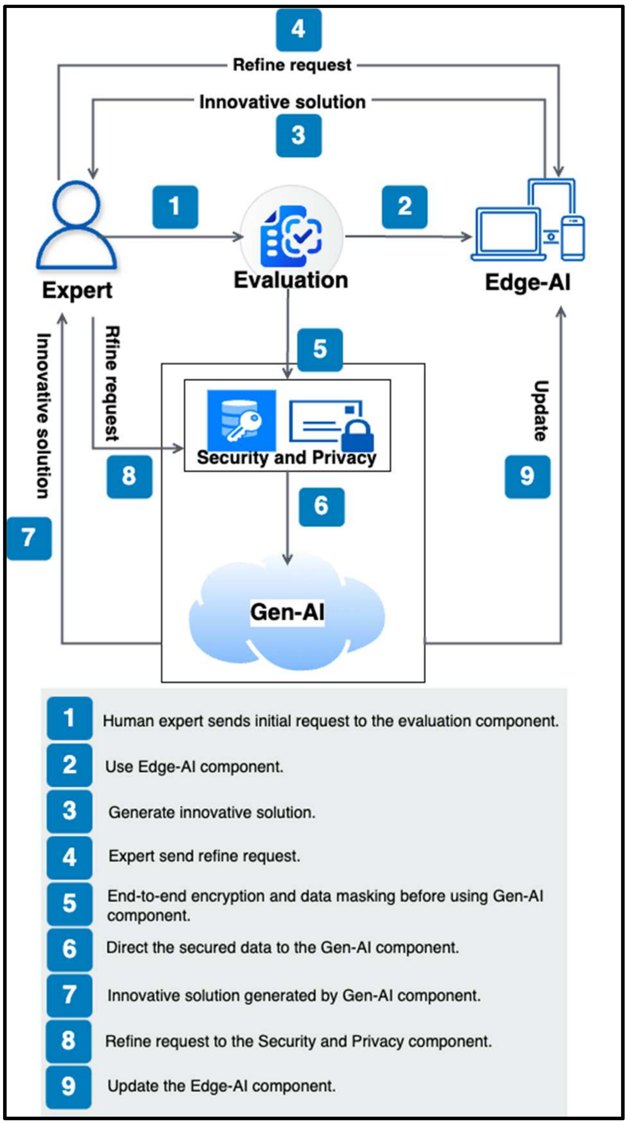
The Gen-Edge-AI framework is structured into two possible solution search paths for highly responsive decision-making. In Path 1, the request is directed to Edge-AI, while Path 2 involves Gen-AI. The decision of which path to follow is determined by the Evaluation Component, based on various routing factors and the human expert's judgment.

The first path begins when the human expert submits an initial request for solution search to the Evaluation component. This component assesses various routing factors, including request complexity, required accuracy, acceptable latency, network resource availability, and data sensitivity. Based on this evaluation, the request is directed to Edge-AI to generate an innovative solution using trained models running on edge devices. By processing the request locally, Gen-Edge-AI minimizes the need for frequent cloud connections, reducing resource consumption [7]. Edge-AI delivers accurate solutions on low-power devices by employing model decompression and optimization techniques [18]. The data are protected as it is processed locally.

The human expert critically reviews the innovative solution based on ethical considerations and human judgment, ensuring that AI-generated outcomes align with required ethical standards. The expert then decides whether to accept the solution or send it back to Edge-AI for refinement. This iterative process continues (step 3 and 4 in Fig. 4) until a satisfactory final solution is achieved or the expert opts to use Gen-AI for further enhancement.

The second path of the solution search involves additional human interaction within the Gen-Edge-AI framework, engaging the Gen-AI component. This path is activated in two scenarios. In the first scenario, the initial request is assessed by the Evaluation component as in the first phase. Then, the human expert makes the decision to proceed with Gen-AI rather than Edge-AI based on that evaluation (step 5 in Fig. 4). In the second scenario, after reviewing the innovative solution for ethical compliance and human judgement, the human expert determines that further fine-tuning by Gen-AI is necessary (step 8 in Fig. *4*). In both scenarios, before involving the Gen-AI component, the Security and Privacy component is used to ensure the secure transmission of the request to the cloud using end-to-end encryption and data masking (step 6 in Fig. 4). Once a satisfactory solution is achieved, it is returned to the human expert and used to update Edge-AI. Optimizing the connection to Gen-AI is essential, as it consumes significant resources [7].

The local processing approach implemented by Edge-AI in the first solution search path reduces the cloud computing load by shifting workloads to edge devices. Additionally, processing data locally minimizes the risk of exposing sensitive data to the cloud. Integrating ethical regulations and



1 Human expert sends initial request to the evaluation component.
2 Use Edge-AI component.
3 Generate innovative solution.
4 Expert send refine request.
5 End-to-end encryption and data masking before using Gen-AI component.
6 Direct the secured data to the Gen-AI component.
7 Innovative solution generated by Gen-AI component.
8 Refine request to the Security and Privacy component.
9 Update the Edge-AI component.

Fig. 4. Gen-Edge-AI Framework

human judgment enriched the proposed framework by incorporating guided human feedback. This approach optimizes bandwidth, enhances privacy, increases security, enables real-time decision-making, and provides personalized experiences. The process flow of this framework is illustrated in Fig. 4 and the components are discussed in detail in the following section.

VI. Gen-Edge-AI Components

In this section, the components of the proposed Gen-Edge-AI framework are discussed to illustrate their individual roles and how they collectively contribute to effective decision-making. Each component plays a unique part in ensuring that the framework operates efficiently, securely, and adaptively across various environments. The integration of these components is critical for balancing the strengths of both Gen-AI and Edge-AI, while incorporating human expertise for real-time, ethical, and high-stakes decision-making. These components are:

A. The human expert component

AI technologies should not compromise sensitive decisions or reduce human control. The human expert component plays a key role in determining the direction of the initial request based on routing factors, as well as deciding whether further refinements are needed for the innovative solution. This human presence ensures that solutions are personalized and beneficial for critical decisions requiring human cognition and intervention (e.g., healthcare emergencies and video surveillance [19]). Therefore, the resulting outcomes are always human-AI guided.

B. The Evaluation component

The Evaluation component is responsible for allocating tasks between Edge-AI and Gen-AI based on routing factors, including request complexity, required accuracy, task urgency (acceptable latency), network bandwidth, data privacy, security requirements, and data sensitivity. It evaluates the sensitivity of the data involved in the expert's request, classifying it into five levels: public, proprietary, private, confidential, and sensitive [20]. This component enables the Gen-Edge-AI framework to operate in a hybrid mode, leveraging both Gen-AI and Edge-AI to tackle more complex problems and deliver innovative solutions effectively.

C. Edge-AI component

The Edge-AI component processes requests that require low latency, minimal bandwidth, and high security for sensitive data. This component reduces cloud dependency by handling tasks locally, resulting in lower computing power consumption and reduced cloud workload. Techniques such as model decompression, quantization, pruning, and knowledge distillation allow for smaller, efficient models to run on edge devices while maintaining high accuracy [18]. Edge-AI creates opportunities for real-time decision-making and improved operational efficiency across various industries.

D. Security and Privacy component

When the Evaluation component directs a request to Gen-AI for further processing, the transmission may introduce privacy risks [13]. To mitigate these risks, the Security and Privacy component ensures that all data are secured through end-to-end encryption and data masking before transmission to the cloud. This approach allows the Gen-Edge-AI framework to operate in a hybrid mode without compromising data security, protecting sensitive information throughout the process.

E. The Gen-AI component

The Gen-AI component handles requests either directly assigned by the Evaluation component or requiring further refinement of an Edge-AI generated solution. Preceded by the Security and Privacy component, this ensures the secure transmission of data to the cloud. The Gen-AI component generates the solution, which is then reviewed by the human expert. A feedback mechanism is used to update the Edge-AI component, enhancing its performance for future similar requests, reducing latency and bandwidth consumption, and minimizing cloud reliance for real-time solutions.

Building on the functionality described in the previous sections, the collaborative integration of all Gen-Edge-AI components makes this framework a highly adaptable solution capable of operating across different modes and environments. The routing factors discussed in the Evaluation component section are critical for understanding how Gen-Edge-AI adaptively allocates requests to deliver reliable, real-time solutions while maintaining data security and resource efficiency. The next section explores these routing factors in detail, highlighting their role in guiding decision-making within the Gen-Edge-AI framework.

VII. Gen-Edge-AI Routing Factors

The proposed Gen-Edge-AI framework balances offline and cloud-based computing by analyzing key routing factors. These factors help determine whether an initial request should be routed to the Edge-AI path (Path 1) or the Gen-AI path on the cloud (Path 2), with the Security and Privacy component applied beforehand to ensure data protection during any cloud interaction. The routing factors are summarized in Table I.

- **Urgency:** The urgency of a request is often highly context-dependent. For example, in real-time scenarios like emergency response, minimizing latency is crucial, whereas in longer-term data analytics, latency may be less significant. A human expert determines the acceptable latency based on the urgency and importance of the specific request. High latency and slow response times negatively affect productivity, collaboration, and user experience. Edge-AI reduces latency by processing data locally, while Gen-AI requires sending data to distant cloud servers, resulting in increased latency. Therefore, if minimizing latency is a priority for an urgent request, the framework uses Edge-AI for a faster, innovative solution. Conversely, when latency is less critical, the request can be processed by Gen-AI.
- **Data sensitivity:** Evaluating the sensitivity of data involves not just technical classifications but also ethical and regulatory considerations. Data can be categorized as low, medium, or highly sensitive data [21]. The implications of data sensitivity differ based on privacy laws, company policies, and ethical considerations, necessitating human oversight to determine the appropriate sensitivity level for handling requests for generating innovative solution. This is the role of the human expert in the proposed framework. Public data, classified as low sensitive, can be processed by the Gen-AI component without significant concern. However, personal and classified data are categorized as medium and highly sensitive [22] and are handled by the Edge-AI component to ensure maximum privacy and security. In cases where medium or highly sensitive data must be processed by the Gen-AI component due to other routing factors, the integrated Security and Privacy component of the proposed framework ensures protection through end-to-end encryption and data masking.

TABLE I. Routing Factors Summary

Routing Factor	Description	AI component selection
Urgency	Determines the criticality of minimizing latency. Real-time, urgent requests require immediate processing.	Edge-AI is used for urgent requests; Gen-AI is used for non-urgent, less time-sensitive requests.
Data Sensitivity	Evaluates the level of data sensitivity (low, medium, or high), considering privacy, regulations, and ethical implications.	Edge-AI handles medium and highly sensitive data for privacy; Gen-AI processes low-sensitivity data or highly protected data.
Security	Assesses the need for enhanced privacy to minimize data exposure.	Edge-AI is used for high security needs; Gen-AI is used with safeguards in place.
Request Intricacy	Considers the complexity of the request, especially when ethical issues are involved	Human Expert assesses whether a request is complex or simple, determining the need for Edge-AI or Gen-AI accordingly.
Solution Accuracy	Evaluates accuracy requirements based on the application domain (e.g., healthcare or content recommendations).	Human Expert refines solutions to ensure accuracy standards are met; Gen-AI or Edge-AI is selected based on these needs.
Network Resource Availability	Examines bandwidth and other network resources to decide on optimal processing.	Edge-AI is preferred to minimize bandwidth use; Gen-AI is used when network resources are sufficient.
Computing Capacity	Determines if high computational power is required, favoring cloud-based processing over limited edge devices.	Gen-AI is used for high computational needs; Edge-AI is used for low to medium power needs.

- **Security:** This factor is highly coupled with data sensitivity. The security architecture of Edge-AI offers enhanced privacy by processing sensitive data locally on the device, whereas Gen-AI requires transmitting data to cloud servers, potentially exposing sensitive data to third parties or vulnerable communication channels. If a request involves sensitive data and security is a top priority, it is handled by the Edge-AI component to minimize risks. If any other factors showed the urgency to use Gen-AI, the Security and Privacy component in the proposed framework is safeguarding the transmitted data.
- **Request intricacy:** Assessing the intricacy of a request involves deep knowledge of multiple layers and context-specific considerations, particularly when ethical factors are involved. It requires careful human judgment to evaluate these complexities. What may appear straightforward in one scenario could be highly intricate in another, depending on the domain and specific requirements. For example, decisions related to healthcare or the environment have significant ethical implications. Human expertise is essential in determining whether a task is simple or complex, especially when considering other dimensions and nuanced factors that can influence the evaluation.
- **Solution accuracy:** Different requests have varying accuracy requirements depending on their intended application. For instance, healthcare applications demand extremely high accuracy, whereas content recommendation systems may tolerate lower accuracy levels. The required level of accuracy also depends on domain-specific knowledge, ethical considerations, and risk factors, which are best assessed by a human expert. In the proposed framework, the human expert is responsible for finalizing and refining innovative solutions created by either Edge-AI or Gen-AI, ensuring they meet the necessary accuracy standards.
- **Network resource availability:** This factor assesses the current capabilities of the network and predicts their impact on response time. Human interpretation and understanding of the request's context are essential for making informed decisions. An example of network resources is bandwidth, which refers to the capacity for inbound and outbound data transfer over a network. Edge-AI requires less bandwidth since data processing occurs locally, whereas Gen-AI involves transferring data to cloud servers, resulting in higher bandwidth demands. When optimizing network resources is a priority, the Edge-AI component is preferred due to its lower

bandwidth requirements, making it a more cost-effective choice.

- **Computing capacity:** Edge-AI has computation capacity boundary due to the constraints of edge devices [15]. In contrast, Gen-AI, hosted on the cloud, provides higher processing capabilities and greater storage capacity, enabling efficient training that can update the Edge-AI component, ultimately resulting in a more sophisticated Gen-Edge-AI model. Therefore, if an initial request requires high computational power, the framework directs it to the Gen-AI component. For requests needing low or medium computing power, the Edge-AI component is chosen.

Even though some of these factors are not inherently subjective, they all require human expertise to ensure decisions are made in an ethical, informed, and contextually-appropriate manner. Human oversight is crucial for evaluating these factors, as they often involve subjective analysis and depend on the nuances of each specific situation. Factors like network resource availability and computing capacity can typically be assessed objectively using measurable parameters such as bandwidth, resource capacity, and system capabilities, but may still require human interpretation for context-specific decisions.

According to the routing factors in the Evaluation component of Gen-Edge-AI framework, the expert assigns weights to each factor based on its importance in the decision-making process. The weights range from 0 to 1 to represent the significance of each factor where 1 is significant and 0 is insignificant. For each factor, score Edge-AI and Gen-AI based on how well they fulfill the request for that particular factor. The scale range from 1 to 5, where 1 means that this AI component is least suitable to the request and 5 means that this AI component is most suitable to the request. Generate the weight column by multiplying the score by the weight for each factor. then, sum the weighted scores to get a final score for Edge-AI and Gen-AI. The expert is now able to determine which option is more suitable for the given request when the final scores for Edge-AI and Gen-AI are compared. Table II presents a hypothetical numerical example of the decision matrix used by an expert in the Evaluation component to make a decision related to the health of the firemen workforce.

TABLE II. ROUTING FACTORS DECISION MATRIX EXAMPLE

Factor	**Weight**	**Edge-AI Score**	**Edge-AI Weighted Score**	**Gen-AI Score**	**Gen-AI Weighted Score**
Urgency	0.25	5	1.25	2	0.5
Data Sensitivity	0.2	5	1.0	3	0.6
Security	0.2	5	1.0	3	0.6
Request Intricacy	0.15	4	0.6	4	0.6
Solution Accuracy	0.1	3	0.3	5	0.5
Network Resource Availability	0.05	4	0.2	3	0.15
Computing Capacity	0.05	2	0.1	5	0.25
Total Score			**4.45**		**3.2**

Based on the scores, Edge-AI is identified as more suitable for this request, with a higher final score of 4.45 compared to Gen-AI's score of 3.2. This decision ensures that factors such as request urgency, data sensitivity, and security are effectively considered in the context of firemen health monitoring.

VIII. POTENTIAL BENEFITS OF GEN-EDGE-AI

The proposed Gen-Edge-AI framework offers significant advantages in scenarios where real-time, innovative solutions are required, while ensuring ethical decision-making through human judgment, minimal latency, and maximum data security and privacy. By integrating Gen-AI and Edge-AI with human expertise, this model effectively addresses the limitations of standalone AI systems, offering a hybrid approach that maximizes the strengths of each component. Key potential benefits of the Gen-Edge-AI framework include:

- **Efficient resource allocation:** Gen-Edge-AI allocates tasks between Gen-AI and Edge-AI based on available resources and routing factors, optimizing computing power, bandwidth usage, and overall efficiency. By preprocessing data locally, the framework reduces the demand on network bandwidth, transmitting only essential information to the cloud-based Gen-AI component. This minimizes data transfer volumes, reduces bottlenecks, and improves operational efficiency—particularly valuable in scenarios with limited network capacity or in remote areas with poor or unavailable internet connectivity.
- **Reliable functionality in limited connectivity:** The Gen-Edge-AI framework can operate effectively in environments with limited or no connectivity, ensuring reliable AI functionality in remote locations, emergency situations, and areas where internet reliability is compromised. This reliability prevents disruptions in critical operations, making it suitable for diverse and challenging environments.
- **Resilience and continuity:** Edge-AI maintains operational continuity even during network disruptions or cloud outages, making it ideal for mission-critical applications. This resilience is vital for sensitive real-life applications that cannot afford interruptions, such as computer vision for object detection, drones inspecting infrastructure for defects, security cameras identifying intruders, and anomaly detection for hospital emergencies.
- **High responsive:** Gen-Edge-AI ensures that data processing is prioritized to occur locally. This significantly minimize the response time, making Gen-Edge-AI promising model for applications where real-time responsiveness is critical. This ability to deliver high responsive solutions is crucial for high-stakes situations that demand rapid and effective decision-making.
- **Enhanced security and privacy**: The proposed framework includes a Security and Privacy component that employs end-to-end encryption and data masking to protect sensitive data. By processing critical data locally whenever possible, Gen-Edge-AI minimizes exposure to privacy risks and ensures compliance with data protection regulations. This approach mitigates risks associated with data transmission, particularly in sensitive environments like public surveillance, where transmitting data to the cloud could pose significant threats.

- **Personalized responses:** Local data processing using Edge-AI allows Gen-Edge-AI to provide highly personalized responses, leveraging localized and context-specific information. This capability enhances the relevance of solutions, leading to improved user satisfaction and more effective outcomes.
- **Ethical and context-aware decision-making**: By incorporating human expertise, Gen-Edge-AI ensures that decision-making is ethically guided and contextually aware, especially in scenarios with moral or ethical considerations, such as healthcare, environmental monitoring, and law enforcement. This human-AI collaboration results in well-rounded solutions that align with ethical standards and address complex, nuanced challenges.
- **Environmental sustainability**: By reducing reliance on cloud data centers and focusing on local processing, Edge-AI helps lower the overall energy consumption of AI operations, contributing to a smaller carbon footprint. This aligns the Gen-Edge-AI framework with sustainable practices, addressing the environmental impact of cloud computing.
- **Adaptive decision-making**: The Gen-Edge-AI framework's adaptability allows it to direct requests based on the most suitable AI component—Edge-AI or Gen-AI—according to operational requirements and routing factors. This ensures that each request is managed in the most efficient and context-appropriate manner, enhancing both accuracy and responsiveness.
- **Scalability and flexibility**: The hybrid nature of Gen-Edge-AI offers improved scalability by distributing workloads between edge devices and the cloud, enabling seamless expansion without overburdening any single component. This flexibility is crucial for supporting diverse operational needs across various environments.

IX. Prospective applications of Gen-Edge-AI

The Gen-Edge-AI framework offers a wide range of potential applications, particularly in scenarios where reliable internet connectivity is unavailable or impractical. By integrating the expertise of human decision-makers with the combined strengths of Gen-AI and Edge-AI, this framework delivers real-time, intelligent solutions while ensuring ethical decision-making, privacy, and security. The human expert plays a critical role in directing requests to the most suitable AI component and ensuring that outcomes align with ethical standards and contextual needs, especially in complex or high-stakes situations. The following are key areas where Gen-Edge-AI framework can be effectively implemented, demonstrating its ability to address challenges in diverse, connectivity-limited environments while ensuring ethical oversight when required, and supporting efficient decision-making:

- Smart phones, smart home devices, and autonomous vehicles: In regional areas with limited connectivity, Gen-Edge-AI can enable devices like smartphones, smart home systems, and autonomous vehicles to function independently. Features such as voice assistants, security monitoring, climate control, and personalized recommendations can operate without relying on continuous cloud connectivity. Drones and autonomous vehicles could navigate, detect obstacles, and make safety-critical decisions. Human expertise would be crucial in guiding these autonomous operations, ensuring that ethical considerations and complex issues are addressed appropriately in real-time, even in challenging environments.
- Wearable and implantable health monitors and healthcare diagnostics: In rural areas, Medical device such as trackers or smartwatches can track vital signals and provide instant and health alerts without requiring cloud connections, making them ideal for rural areas without the need to cloud connection. In under-resourced clinics, Gen-Edge-AI can assist healthcare professionals by processing medical images and predicting health outcomes locally. The healthcare provider (the human expert) can then make the final decision for ethical patient care or complex cases that require human judgment.
- Disaster management and emergency response: In natural disaster scenarios where internet connectivity is unreliable, Gen-Edge-AI can enable drones and autonomous vehicles to conduct search and rescue operations or deliver critical supplies to affected areas. Crisis management professionals (human experts) provide ethical oversight and contextual decision-making, particularly for sensitive tasks such as prioritizing victims in critical situations.
- Agricultural monitoring: IoT devices deployed in farmlands in remote areas collect data on soil quality, moisture levels, and crop health. Gen-Edge-AI can process this data locally to provide farmers with immediate insights, enabling them to intervene in more complex situations to ensure ethical practices in land and water resource management.
- Wildlife monitoring and conservation: Gen-Edge-AI can be deployed to monitor wildlife habitats and can analyze data locally to identify threats or changes in the environment. Human conservation experts will then be able to use these insights to take timely, ethically informed action to protect endangered species.
- Industrial automation and manufacturing: In smart factories, Gen-Edge-AI can control production lines and automates quality checks by processing sensor data locally for real-time decision-making. When anomalies arise indicating potential ethical issues or risks, human experts intervene to make context-driven decisions, ensuring both production safety and efficiency, while effectively addressing any ethical or risk-related concerns.
- Supply chain and logistics: In isolated regions, Gen-Edge-AI can be used in logistics to optimize delivery routes and inventory levels by processing data locally. This minimizes the need for constant internet access and allows human experts to make context-specific adjustments when ethical concerns or unusual situations arise.
- Public safety and law enforcement: Cameras equipped with Gen-Edge-AI can detect incidents or suspicious activities without relying on a cloud connection. Meanwhile, human operators can evaluate the AI's findings, ensuring that actions taken are ethically

sound and respect individual rights and privacy regulations.

- Defense and military: For military use cases, where communication infrastructure might be compromised and more data are classified, Gen-Edge-AI can handle immediate, critical tasks like threat detection or navigation. In Gen-Edge-AI, human military operators play a key role in making the final decisions, ensuring that the procedure operates ethically and in line with strategic goals.
- Space edge computing: Space environments face challenges related to size, power constraints, and radiation exposure, which limit the capabilities of onboard processing systems. Gen-Edge-AI is potentially beneficial to be used in domain-specific architectures with specialized hardware accelerators to address these challenges effectively with the support of the human expert.

These examples illustrate the broad applicability of the Gen-Edge-AI framework and shows its versatility in environments with connectivity limitations. The framework values the human judgment in ensuring ethical and contextually appropriate decision-making.

X. Conclusion and Future work

In conclusion, the Gen-Edge-AI framework is ideal for developing models where rapid response times are critical for situations involving ethical and contextual human judgement. The hybrid nature of Gen-Edge-AI allows it to optimize network resources, reduce data transmission risks, and minimize bandwidth usage, while keeping human oversight central to handling ethical and high-stakes real-time decisions. The prospective applications of Gen-Edge-AI in various fields, from healthcare and emergency response to creative industries and smart infrastructure, present a promising opportunity for evaluating its return on investment and operational resilience. Future research is recommended to focus on expanding the framework's capabilities, such as integrating a dynamic decision support system to allocate requests more efficiently and incorporating dynamic load balancing. Real-world applications will showcase the effectiveness of the Gen-Edge-AI framework. As Gen-AI and Edge-AI continue to reshape human-AI interaction, the proposed Gen-Edge-AI provides a powerful blend of technology and human judgment that ensures solutions are both innovative and ethically grounded.

References

[1] E. Veitch and O. Andreas Alsos, "A systematic review of human-AI interaction in autonomous ship systems," *Saf Sci*, vol. 152, p. 105778, 2022, doi: 10.1016/j.ssci.2022.105778.

[2] Statista, "Share of internet users in the United States who have tried generative artificial intelligence (AI) as of the second quarter of 2023, by generation," https://www.statista.com/statistics/1489374/us-generative-ai-usage-by-generation/.

[3] "This person doesn't exist," https://this-person-does-not-exist.com/en#google_vignette. Accessed: Oct. 04, 2024. [Online]. Available: https://this-person-does-not-exist.com/en#google_vignette

[4] Dell Technologies Inc, "The impact of AI on edge computing," *CIO*, Apr. 2024, Accessed: Oct. 02, 2024. [Online]. Available: https://www.proquest.com/trade-journals/impact-ai-on-edge-computing/docview/3049738771/se-2?accountid=14757

[5] P. Gupta, B. Ding, C. Guan, and D. Ding, "Generative AI: A systematic review using topic modelling techniques," *Data Inf Manag*, vol. 8, no. 2, p. 100066, 2024, doi: https://doi.org/10.1016/j.dim.2024.100066.

[6] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning," *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023.

[7] Sanjeev. Pulapaka, Srinath. Godavarthi, and Sherry. Ding, *Empowering the Public Sector with Generative AI : From Strategy and Design to Real-World Applications* , 1st ed. 2024. Berkeley, CA: Apress, 2024.

[8] K. Sengupta and P. R. Srivastava, "HRNET: AI-on-Edge for Mask Detection and Social Distancing Calculation," *SN Comput Sci*, vol. 3, no. 2, p. 157, 2022.

[9] D. Uprety, D. Zhu, and H. (Jack) West, "ChatGPT—A promising generative AI tool and its implications for cancer care," *Cancer*, vol. 129, no. 15, pp. 2284–2289, 2023.

[10] Yasir Raza, "Cutting-Edge AI Tools: An Up-to-Date Dataset - Explore the Latest Advancements in AI: A Comprehensive Dataset of Emerging Tools," https://www.kaggle.com/datasets/yasirabdaali/740-ai-tools-for-everyone.

[11] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," *arXiv preprint arXiv:1802.05668*, 2018.

[12] P. Zhang, G. Zeng, T. Wang, and W. Lu, "Tinyllama: An open-source small language model," *arXiv preprint arXiv:2401.02385*, 2024.

[13] E. Villar-Rodriguez, M. A. Pérez, A. I. Torre-Bastida, C. R. Senderos, and J. López-de-Armentia, "Edge intelligence secure frameworks: Current state and future challenges," *Comput Secur*, vol. 130, p. 103278, 2023.

[14] Z. Zhang, F. Li, C. Lin, S. Wen, X. Liu, and J. Liu, "Choosing Appropriate AI-enabled Edge Devices, Not the Costly Ones," in *2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS)*, IEEE, 2021, pp. 201–208.

[15] A. Jevremovic, Z. Kostic, and D. Perakovic, "Energy-Efficient Edge Intelligence: A Comparative Analysis of AIoT Technologies," *Mobile networks and applications*, 2023.

[16] E. Veitch and O. Andreas Alsos, "A systematic review of human-AI interaction in autonomous ship systems," *Saf Sci*, vol. 152, p. 105778, 2022, doi: https://doi.org/10.1016/j.ssci.2022.105778.

[17] L. Boussioux, J. N. Lane, M. Zhang, V. Jacimovic, and K. R. Lakhani, "The Crowdless Future? Generative AI and Creative Problem-Solving," *Organization Science*, vol. 35, no. 5, pp. 1589–1607, 2024, doi: 10.1287/orsc.2023.18430.

[18] C. Surianarayanan, J. J. Lawrence, P. R. Chelliah, E. Prakash, and C. Hewage, "A Survey on Optimization Techniques for Edge Artificial Intelligence (AI)," *Sensors*, vol. 23, no. 3, p. 1279, Jan. 2023, doi: 10.3390/s23031279.

[19] A. N. Gohar, S. A. Abdelmawgoud, and M. S. Farhan, "A Patient-Centric Healthcare Framework Reference Architecture for Better Semantic Interoperability Based on Blockchain, Cloud, and IoT," *IEEE Access*, vol. 10, pp. 92137–92157, 2022, doi: 10.1109/ACCESS.2022.3202902.

[20] R. Bragg, " CISSP Security Management and Practices," in *CISSP Training Guide*, Pearson IT Certification., 2002, ch. 3.

[21] J. M. M. Rumbold and B. K. Pierscionek, "What Are Data? A Categorization of the Data Sensitivity Spectrum," *Big Data Research*, vol. 12, pp. 49–59, 2018, doi: 10.1016/j.bdr.2017.11.001.

[22] F. Jin, S. Wu, X. Liu, H. Su, and M. Tian, "Detection of Unstructured Sensitive Data Based on a Pre-Trained Model and Lattice Transformer," in *2024 7th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2024, pp. 180–185. doi: 10.1109/ICAIBD62003.2024.10604568.